Artificial Intelligence
Assignment 4

Nakul Thureja
2020528

Data Manipulation:
I have used the label encoders to convert string type into numerical values. This is required to make the data fit for running in an ML model.

I have also removed a few columns which I felt were acting like noise and weren't helping in data prediction.
Columns removed:

- 'In a Realtionship?'
- 'interested in games'
- 'Interested Type of Books'
- 'Gentle or Tuff behaviour?'

Data Training:

1. Raw Data
   When I trained the data with the raw model after performing the steps given above. I got an accuracy of around 5%. The classification report also shows very low accuracies.

2. Modified Data
   I modified the data by grouping the Suggested Job Role under 6 categories. Then I ran the MLP model again and got an accuracy of around 45%.

3. Correlated Data from Assignment 1
   For correlating the data with assignment 1, I took the interested subject and interested careers fields in the data. Further, I applied the rules from my 1st assignment and generated a new Suggested Job Role column. Then I ran the MLP model again and got an accuracy of around 98%, this could be possible since the subjects and careers are correlated now and we are getting better results from the MLP model.

Parameters of MLP classifier Model:

- "relu" activation function
- "adam" solver
- "(128, 64, 32)" Hidden layers
- "500" Maximum Iterations

Results:

1. Raw Data

```
Training accuracy on raw data: 0.0566875
Testing accuracy on raw data: 0.0515
```

	precision	recall	f1-score	support
0	0.00	0.00	0.00	105
1	0.00	0.00	0.00	98
2	0.00	0.00	0.00	112
3	0.00	0.00	0.00	130
4	0.00	0.00	0.00	116
5	0.00	0.00	0.00	112
6	0.00	0.00	0.00	108
7	0.00	0.00	0.00	116
8	0.00	0.00	0.00	102
9	0.00	0.00	0.00	124
10	0.00	0.00	0.00	118
11	0.00	0.00	0.00	110
12	0.00	0.00	0.00	120
13	0.00	0.00	0.00	128
14	0.00	0.00	0.00	107
15	0.00	0.00	0.00	118
16	0.06	1.00	0.11	224
17	0.00	0.00	0.00	121
18	0.00	0.00	0.00	110
19	0.00	0.00	0.00	126
20	0.00	0.00	0.00	131
21	0.00	0.00	0.00	93
22	0.00	0.00	0.00	104
23	0.00	0.00	0.00	118
24	0.00	0.00	0.00	121
25	0.00	0.00	0.00	116
26	0.00	0.00	0.00	132
27	0.00	0.00	0.00	102
28	0.00	0.00	0.00	124
29	0.00	0.00	0.00	90
30	0.00	0.00	0.00	104
31	0.00	0.00	0.00	124
32	0.00	0.00	0.00	115
33	0.00	0.00	0.00	121
accuracy			0.06	4000
macro avg	0.00	0.03	0.00	4000
weighted avg	0.00	0.06	0.01	4000

2. Modified Data

80-20 split

```
Training Accuracy with Modified Data (80-20 split):  0.4616875
Testing Accuracy with Modified Data (80-20 split):  0.4505
```

Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	611
1	0.00	0.00	0.00	1123
2	0.00	0.00	0.00	122
3	0.45	1.00	0.62	1803
4	0.00	0.00	0.00	93
5	0.00	0.00	0.00	248
accuracy			0.45	4000
macro avg	0.08	0.17	0.10	4000
weighted avg	0.20	0.45	0.28	4000

Confusion Matrix:

```
[[   0    0    0  611    0    0]
 [   0    0    0 1123    0    0]
 [   0    0    0  122    0    0]
 [   0    1    0 1802    0    0]
 [   0    0    0   93    0    0]
 [   0    0    0  248    0    0]]
```

70-30 split

```
Training Accuracy with Modified Data (70-30 split):  0.46185714285714285
Testing Accuracy with Modified Data (70-30 split):  0.4523333333333333
```

Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	852
1	0.14	0.00	0.00	1719
2	0.00	0.00	0.00	179
3	0.45	1.00	0.62	2720
4	0.00	0.00	0.00	168
5	0.00	0.00	0.00	362
accuracy			0.45	6000
macro avg	0.10	0.17	0.10	6000
weighted avg	0.25	0.45	0.28	6000

Confusion Matrix:

```
[[   0    2    0  850    0    0]
 [   0    2    0 1717    0    0]
 [   0    1    0  178    0    0]
 [   1    7    0 2712    0    0]
 [   0    1    0  167    0    0]
 [   0    1    0  361    0    0]]
```

60-40 split

```
Training Accuracy with Modified Data (60-40 split):  0.4555
Testing Accuracy with Modified Data (60-40 split):  0.465375
```

Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	1119
1	0.00	0.00	0.00	2221
2	0.00	0.00	0.00	214
3	0.47	1.00	0.64	3724
4	0.00	0.00	0.00	246
5	0.00	0.00	0.00	476
accuracy			0.47	8000
macro avg	0.08	0.17	0.11	8000
weighted avg	0.22	0.47	0.30	8000

Confusion Matrix:

```
[[   0    0    0 1119    0    0]
 [   1    0    0 2220    0    0]
 [   0    0    0  214    0    0]
 [   1    0    0 3723    0    0]
 [   0    0    0  246    0    0]
 [   0    0    0  476    0    0]]
```

3. Correlated Data from Assignment 1

80-20 split

```
Training Accuracy of Correlated Data from Assigment 1 (80-20 split):  0.9929375
Testing Accuracy of Correlated Data from Assigment 1 (80-20 split):  0.99
```

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	787
1	0.98	0.99	0.99	496
2	0.99	0.98	0.99	186
3	0.99	0.99	0.99	1373
4	0.97	0.99	0.97	336
5	0.99	0.99	0.99	822
accuracy			0.99	4000
macro avg	0.99	0.99	0.99	4000
weighted avg	0.99	0.99	0.99	4000

Confusion Matrix:

```
[[ 787    0    0    0    0    0]
 [   0  492    1    3    0    0]
 [   0    2  182    0    0    2]
 [   2    7    0 1354    6    4]
 [   0    0    0    4  331    1]
 [   0    0    0    2    6  814]]
```

70-30 split

```
Training Accuracy of Correlated Data from Assigment 1 (70-30 split):  0.9928571428571429
Testing Accuracy of Correlated Data from Assigment 1 (70-30 split):  0.9891666666666666
```

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1213
1	0.98	0.99	0.99	692
2	0.98	1.00	0.99	297
3	1.00	0.99	0.99	2067
4	0.96	0.97	0.97	519
5	0.99	0.98	0.98	1212
accuracy			0.99	6000
macro avg	0.98	0.99	0.99	6000
weighted avg	0.99	0.99	0.99	6000

Confusion Matrix:

```
[[1213    0    0    0    0    0]
 [   0  686    4    2    0    0]
 [   0    0  296    0    0    1]
 [   0   11    0 2042    3   11]
 [   0    0    0    7  506    6]
 [   0    0    2    0   18 1192]]
```

60-40 split

```
Training Accuracy of Correlated Data from Assigment 1 (60-40 split):  0.9909166666666667
Testing Accuracy of Correlated Data from Assigment 1 (60-40 split):  0.9865
```

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1558
1	0.98	0.99	0.98	917
2	0.96	0.97	0.96	418
3	0.98	0.99	0.99	2832
4	0.99	0.98	0.98	655
5	0.99	0.97	0.98	1620
accuracy			0.99	8000
macro avg	0.98	0.98	0.98	8000
weighted avg	0.99	0.99	0.99	8000

Confusion Matrix:

```
[[1558    0    0    0    0    0]
 [   0  906    0   11    0    0]
 [   0    2  404    0    0   12]
 [   0   16    0 2811    1    4]
 [   0    0    0   12  639    4]
 [   0    0   19   22    5 1574]]
```